基于标签分布学习的轻量级人脸表情识别研究 *

刘 劲 [1]，罗晓曙 [1†]，徐照兴 [2]

(1. 广西师范大学 电子工程学院，广西 桂林 541000; 2. 江西服装学院 大数据学院，南昌 330000)

摘 要：针对轻量级网络在复杂环境下对面部表情的特征提取不够充分、泛化能力不足以及单标签数据集无法有效描述复杂情感倾向所带来的歧义表情等问题，提出了一种结合改进 ShuffleNet 与标签分布学习的人脸表情识别方法。在不大量增加计算复杂度的前提下，为了避免模型的过拟合，设计了新的输出模块对 ShuffleNet 模型进行改进；为了增强模型对人脸表情图像重要局部细节特征的提取能力，设计了并行的深度卷积残差模块，实现了局部与全局特征的融合。为了减少由歧义表情对识别性能所带来的不利影响，通过标签分布学习方法，在不引入额外信息的前提下，充分利用数据集原本信息生成标签分布，并重新训练改进后的 ShuffleNet 模型。实验结果表明，在 RAF-DB、AffectNet-7 和 AffectNet-8 数据集上分别达到了 87.15%、62.05%和 58.49%的准确率，同时参数量和计算量均保持在较低水平，利于其在实际生产中应用。

关键词：人脸表情识别；轻量化；标签分布学习；歧义表情；深度可分离卷积

中图分类号：TP391.4 doi: 10.19734/j.issn.1001-3695.2021.12.0697

Research on lightweight facial expression recognition based on label distribution learning

Liu Jin[1], Luo Xiaoshu[1†], Xu Zhaoxing[2]

(1. *School of Electronic Engineering*, *Guangxi Normal University*, *Guilin Guangxi* 541000, *China*; 2. *School of Big Data*, *Jiangxi Institute of Clothing University*, *Nanchang Jiangxi* 330000, *China*)

Abstract: Aiming at the problems of insufficient facial expression feature extraction in complex environments, insufficient generalization ability, and single-label data sets that cannot effectively describe the ambiguous expressions caused by complex emotional tendencies, this paper proposed a facial expression recognition method combining improved ShuffleNet and label distribution learning. On the premise of not greatly increasing the computational complexity, to avoid over-fitting of the model, designed a new output module to improve the ShuffleNet; to enhance the model's ability to extract important local details of facial expression images, designed a parallel depthwise convolution residual module to realize the fusion of local and global features. In order to reduce the negative impact of ambiguous expressions on recognition performance, used the label distribution learning method to make full use of the original information of the data set to generate the label distribution without introducing additional information and retrain the improved ShuffleNet model. The experimental results show that the accuracy rates of 87.15%, 62.05% and 58.49% are achieved on the facial expression data sets RAF-DB, AffectNet-7 and AffectNet-8, at the same time, the number of parameters and FLOPs are kept at a low level, which is conducive to its application in actual production.

Key words: facial expression recognition; lightweight; label distribution learning; ambiguous expressions; depthwise separable convolution

0 引言

自古，“观色”是全面分析人物心理活动的重要依据。在《论语·颜渊》中更是有：“夫达也者，质直而好义，察言而观色，虑以下人”。通过识别人脸表情来以观其色，可以为出现在场景中的人物提供辅助的结构化信息。因此，人脸表情识别(facial expression recognition，FER)在情感计算、人机交互、驾驶员疲劳检测、教学效果评价等众多领域有着广泛的应用[1,2]。1978 年，Ekman 等人[3]发表的跨文化研究中首次定义了六种基本面部表情：高兴、伤心、生气、害怕、厌恶和惊讶，这些基本情绪可以被不同文化背景的人感知、认同和理解。

近年来，随着深度学习在计算机视觉领域的飞速发展，它也被成功地应用在人脸表情识别领域，并取得了良好的进展。深度学习技术在使表情识别准确率提升的同时，也会导致参数量和 FLOPs 的急剧增加，虽然更大更深的网络模型效果更好，但是模型运行时对所需要的硬件配置要求也越高。而在实际生产与应用环境中，设备的配置水平往往受到成本限制，过高的配置需求不利于模型的实际应用。因此，在人脸表情识别领域除了在提高识别准确率的同时，也应考虑如何压缩模型的计算开销，使模型能够在性能较低的小型嵌入式设备上正常运行。

1980 年，心理学家 Plutchik 等人[4]的研究表明：人类的大多数情绪都是由基本面部表情组成。在现实世界中，某一静态的人脸表情图像往往由不同强度的基本情绪组成，有复杂的情感意图，但表情图像却只对应一个标签。由于这种歧义表情的存在，这使得表情识别的效果严重受限，通过标签分布学习(Lable Distribution Learning，LDL)来解决单标签无法有效描述复杂情感倾向的问题，可以进一步提高 FER 模型的识别性能。此外，标签分布学习还可以缓解由数据集标注

收稿日期：2021-12-28；**修回日期：**2022-03-07 **基金项目：**广西人文社会科学发展研究中心科学研究工程•创新创业专项(重大委托项目)(ZDCXCY01)；广西自然科学基金资助项目(2018GXNSFAA281351)；广西科技重大专项(桂科 AA18118004)

作者简介：刘劲(1997-)，男，湖北天门人，硕士研究生，主要研究方向为计算机视觉；罗晓曙(1961-)，男(通信作者)，湖北应城人，教授，博士，主要研究方向为人工智能技术与应用(lxs@mailbox.gxnu.edu.cn)；徐照兴(1979-)，男，江西抚顺人，教授，硕士，主要研究方向为计算机应用软件开发.

者的主观性和表情图像的模糊性造成的噪声问题[5]。

针对上述两个问题，本文在轻量级网络 ShuffleNet 的基础上构建深度可分离卷积残差模块，在不大量增加计算开销的同时，可以更好的提取人脸表情图像中眼睛、嘴巴等关键细节部位的特征。在训练时，利用 LDL 方法来生成标签分布，这有利于提高模型对不同表情的判别能力，从而提出了基于标签分布学习的轻量级人脸表情识别(lightweight facial expression recognition based on label distribution learning，LFER-LDL)，本文方法在 RAF-DB[6]和 AffectNet-7[7]数据集上进行实验验证，实验结果表明所提方法在保持较低计算开销的情况下，较近期提出的一些表情识别方法有较好的识别性能提升。

1 本文方法

本文提出的人脸表情识别模型主要由改进的 ShuffleNet 网络和标签分布学习(LDL)两部分组成组成，网络结构如图 1 所示。本文的骨干网络为 ShuffleNet-V2[8]，该模型由 Conv1、Stage2、Stage3、Stage4、Conv5 组成。为了避免过拟合，使模型具有更好的鲁棒性，本文设计了新的输出模块来代替原始网络的全连接层。为了增强网络对局部细节特征的提取能力，在不大量增加额外计算开销的前提下，设计了并行深度卷积残差模块(Parallel Depthwise convolution Residual module，PDWRes)。根据 Plutchik 等人[4]的研究，为了减少歧义表情带来的不利影响，在不使用额外信息量的前提下，利用数据集本身来生成标签分布(图 1 的右分支)。改进后的 ShuffleNet 网络(图 1 的左分支)从 Conv1~Conv5 层的输出通道数分别为 29、116、232、464、1024，最后通过 Softmax 层，得到七分类人脸表情识别输出。

1.1 改进的 ShuffleNet 模型

深层的卷积神经网络(convolutional neural network，CNN)如 ResNet 和 VGG 等可以取得较高的表情图像分类准确率，但模型的计算复杂度也随之提升，过于复杂的网络无法满足嵌入式设备场景的需求，一些移动端设备也需要又快又准的小模型。为了满足这些需求，Ma N 等人[8]提出了轻量级神经网络 ShuffleNet-V2，它可以很好的平衡识别准确率和计算速度的关系。在 ShuffleNet-V2 Unit 中主要使用了 1×1 点卷积(pointwise convolution，PWConv)和深度卷积(depthwise convolution，DWConv)，并对不同特征组内的通道信息进行 Channel Shuffle 操作，实现不同组之间的信息融合。

Lin M 等人[9]的研究表明：在 CNN 模型中参数占比最大的是全连接层。虽然全连接层可以压缩特征图(feature map)的维度并输入到 softmax 层，最终得到七分类人脸表情图像，但这会造成过拟合，不利于增强模型的泛化能力。为此，本文设计了改进的输出模块来替换骨干网络 ShuffleNet-V2 中的全连接层输出模块，改进输出模块如图 2 所示。

改进输出模块主要由改进的深度可分离卷积组成，这与骨干网络中的点卷积和深度卷积类似。深度可分离卷积的卷积层通道相关性和空间相关性是可解耦合性的[10]，相较于普通卷积，深度可分离卷积模块可以在进一步提取人脸表情特征的同时不引入大量的参数。当大小为 $d\times d$ 的卷积核作用在大小为 $H\times W$ 的输入特征矩阵上时，令输入、输出的通道数分别为 C 和 n，可得普通卷积计算参数量为 $H\times W\times C\times(d\times d\times n)$，而深度可分离卷积的计算参数量为 $H\times W\times C\times(d\times d+n)$。因此，深度可分离卷积的参数量仅为标准卷积的 $\frac{1}{d^2}+\frac{1}{n}$ 倍。

为了防止梯度弥散，增强模型的非线性能力，减少过拟合，深度可分离卷积后均使用了 ReLu 激活函数，虽然其在反向传播时速度较快，但对于输入不大于 0 的神经元将会被抑制，导致权重无法更新，这会影响整个模型的最终表达。本文对深度可分离卷积模块进行改进，将 ReLu 激活函数替换为 Mish 激活函数。Mish 激活函数公式为

$$\mathrm{Mish}=x*\tanh(\ln(1+e^{x})) \tag{1}$$

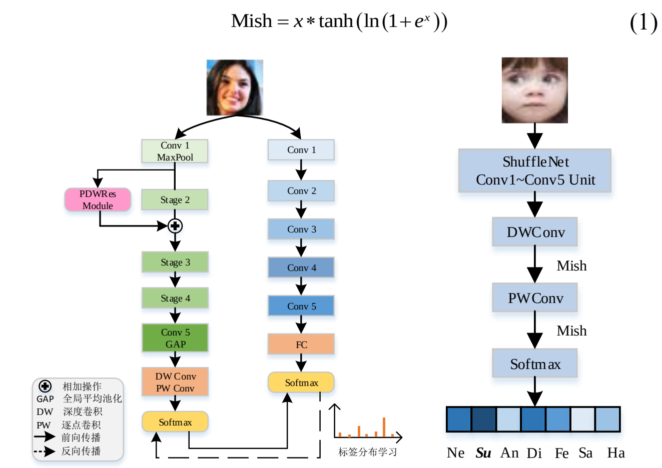


图 1 本文的表情识别网络结构 图 2 改进输出模块流程图

Fig. 1 The expression recognition network structure of this article Fig. 2 Improve the output module flow chart

Mish 激活函数曲线图如图 3，它对负值保留了一定的梯度流，而不像 ReLu 中的硬零边界，这利于特征信息的流动。此外，Mish 曲线上的每一点都是平滑的，这将允许更好的信息深入神经网络，从而取得更好的识别准确率和泛化性。

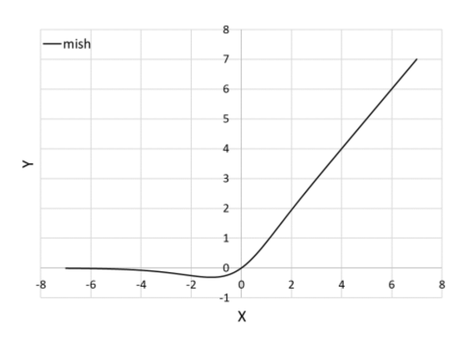


图 3 Mish 激活函数曲线图

Fig. 3 Mish activation function graph

1.2 并行深度卷积残差模块的设计

人脸表情识别往往与局部细节特征有关，例如眉毛、眼睛、嘴巴等部位可以更容易地表现出不同的情绪，人眼在识别表情时也往往关注这些区域。因此，为了使网络可以有效的学习局部细节特征，本文设计了并行的深度卷积残差模块(PDWRes)，通过对局部区域的特征提取，并以残差结构的形式补全到骨干网络中，实现了局部与全局特征的融合，使网络更加关注人脸表情图像中的重要性特征，PDWRes 模块结构如图 4。

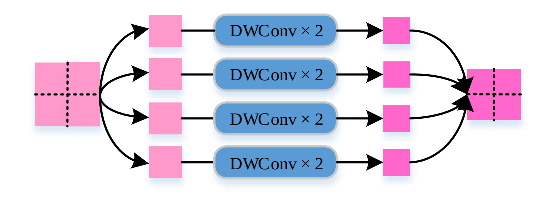


图 4 PDWRes 模块结构图

Fig. 4 Pdwres module structure diagram

对于输入大小为 224×224 的 RGB 人脸表情图像，在通过底层 Conv1 之后得到特征图 $F_1\in\mathbb{R}^{H\times W\times c}$，其中 $H=W=56, c=29$。受到近期 Transformer 模型[11]的启发，

将特征图 F_1 进行水平、垂直方向二等分，得到四块人脸表情的区域特征图 $F_1^k\in\mathbb{R}^{h\times w\times c}$，其中 $h=w=28$，$k=\{1,2,3,4\}$。再

对每小块特征图依次经过两次 3×3 DWConv 操作，得到人脸不同区域的细节特征图 $F_{PDWRes}^{k} \in \mathbb{R}^{h/2 \times w/2 \times C}$，其中 C=116。如 1.1 节所述，为了避免引入大量计算参数，这里仅使用深度卷积提取特征。为了加快模型的收敛速度，在每一次深度卷积之后均使用了批量归一化(batch normalization，BN)，为了增强模型的稀疏性，减少冗余度，在 BN 后同时使用 ReLu6 激活函数，ReLu6 定义如下：

$$\mathrm{ReLu6}(x)=\min(\max(0,x),6)=\begin{cases}6, & x\geq 6\\ x, & 0<x<6\\ 0, & \text{其他}\end{cases} \tag{2}$$

ReLu6 激活函数将 ReLu 函数线性部分的上限设为 6，这有利于在低精度的移动端设备上实现更好的数值分辨率，增强模型的稳定性。

最后，将四块区域特征图 F_{PDWRes}^{k} 沿着水平和垂直方向进行拼接，可得完整的局部特征图 $F_{PDWRes} \in \mathbb{R}^{h \times w \times C}$，并将 F_{PDWRes} 补充到经过 Stage2 之后的全局特征 F_{Stage2} 中，可得全局与局部特征融合表达式为

$$F = F_{Stage2} + F_{PDWRes} \tag{3}$$

由于随着网络深度的加深，特征图将越来越小，这将不利于 PDWRes 模块进行局部特征提取。因此，为了尽可能地减少对模型引入额外的计算量，本文只在 Stage2 阶段使用了 PDWRes 模块。

1.3　标签分布学习

人脸表情图像的标注往往需要大量的人力物力，且情感分布难以获得，这会造成歧义表情，不利于表情图像的分类。为了弥补表情分类时单标签信息量的不足，本文使用了标签分布学习的方法来生成表情分布，如图 1 的右分支，其骨干网络为 ResNet-50，将不同的单标签人脸表情数据集在该标签分布学习方法上进行预训练，收集人脸表情数据集整体的分布，再将生成的数据标签分布重新训练改进后的 ShuffleNet 网络。

给定一张人脸表情图像 x，其标签 $Y \in \{1,2,\ldots,i\}$，其中 i 表示表情图像的类别数，标签分布学习将会收集数据集中表情图像的分布 $P \in (0,1)$。通过 ResNet-50 的全连接层(FC)之后可得 $\omega^{\mathrm{T}} x$，标签分布学习最后以 Softmax 层作为输出，Softmax 计算表情图像 x 属于类别 i 的条件概率为

$$P(Y=i\mid x)=Softmax(w_i^{\top}x)=\frac{exp(w_i^{\top}x)}{\sum_{j=1}^{i} exp(w_j^{\top}x)} \tag{4}$$

其中，ω_i 是第 i 类的权重向量，j 表示总类别数。LDL 的输出结果是输入表情图像 x 属于 7 种不同表情的概率，这些概率之和为 1。

为了利于梯度的反向传播，本文使用 KL 散度来度量改进后 ShuffleNet 模型的预测输出与 LDL 得到标签分布之间的差异。KL 散度是非负的，这满足深度学习梯度下降法特性，但由于其具有非对称性，本文将 LDL 得到的标签分布作为数据的真实分布 $P(x)$，改进后 ShuffleNet 模型的输出作为拟合分布 $Q(x)$，由此，样本数量为 N 的 KL 散度可写为

$$\mathcal{L}_1=\frac{1}{N\times i}\sum_{n=1}^{N}\sum_{i=1}^{i}P(x_i^n)log\left(\frac{P(x_i^n)}{Q(x_i^n)}\right) \tag{5}$$

标签分布学习及 KL 散度只在训练时使用，用于帮助改进后 ShuffleNet 网络更好的学习数据集中人脸表情的分布与判别。在测试时，仅根据改进 ShuffleNet 模型 softmax 层输出概率的最大值作为网络的输出。

在测试阶段，使用 combo loss 作为损失函数，它由改进的交叉熵(CE loss)与 dice loss 的加权和构成。为了控制对不同数据集中假阳性(false positive，FP)和假阴性(false negative，FN)的正则化程度，纠正网络的学习，将二进制交叉熵推广到多分类问题，其输出是多个二进制交叉熵的平均值。Dice Loss 主要用于处理数据集中类别不平衡问题，减小模型在易分类表情上的过拟合。combo loss 可以写为

$$\mathcal{L}_2=\alpha\left(-\frac{1}{N}\sum_{i=1}^{N}\beta(t_i\ln p_i)+(1-\beta)[(1-t_i)\ln(1-p_i)]\right)-(1-\alpha)\left(\frac{2\sum_{i=1}^{N}t_i p_i+\varepsilon}{\sum_{i=1}^{N}t_i+\sum_{i=1}^{N}p_i+\varepsilon}\right) \tag{6}$$

其中，t_i 和 p_i 分别表示真实值与预测值。超参数 α 平衡 combo loss 与改进交叉熵的权重。超参数 β 控制对 FP 与 FN 的正则化程度，实验时根据不同数据集调整。为了避免分母为 0，实验时 ε 取 1 进行平滑。

2　实验

2.1　数据集介绍

本文实验在大规模人脸表情数据集 RAF-DB[6] 和 AffectNet[7] 数据集上进行实验评估，其中 RAF-DB 和 AffectNet-7 均为 7 种类别的表情标签：悲伤、惊讶、厌恶、恐惧、快乐、愤怒、中立，AffectNet-8 数据集在此基础上增加了蔑视的表情，有 8 种类别的表情标签。

RAF-DB 数据集是真实世界人脸情感数据库(Real-world Affective Faces Database)，共有七分类表情图像 15339 张，每张图像均由 40 人独立标注，分为 12271 张训练集和 3068 张测试集。这些表情图像存在着遮挡、姿势、光照条件等不同方面的影响，具有较大的差异性与实际应用价值。

AffectNet 是迄今为止最大的人脸表情数据集，包含超过 100 万张来自互联网的面部图片，这些图片通过不同的搜索引擎检索情感标签获得，其中大约一半(44 万)的图像被标注了 11 种表情类别。本文使用 AffectNet 数据集中手动标记的 29 万张表情图像用作训练集，在 AffectNet-7 中有 3500 张测试图像，在 AffectNet-8 中有 4000 张测试图像。图 5 展示了 RAF-DB 和 AffectNet-7 数据集上的表情图像样例。



图 5　数据集图像样例

Fig. 5　Sample dataset image

2.2　实验环境与数据预处理

本文实验均在 Ubuntu 16.04 系统下完成，基于深度学习框架 PyTorch 1.1 和解释器 Python 3.7 实现，硬件环境：CPU 为 E5-2637 v4，GPU 为 NVIDIA GeForce GTX 1080Ti，显存大小为 11GB，加速库为 CUDA 10.2。

在真实场景采集的 RAF-DB 和 AffectNet 数据集里，表情图像中人脸的大小、角度、姿势各有不同，这不利于模型的学习，因此均使用了 Retinaface[12]进行人脸检测和对齐。为了优化模型的学习效率，本文方法在 MS-Celeb-1M 人脸数据集上进行预训练。为了避免过拟合，将 RAF-DB 和 AffectNet 数据集所有表情图像的大小均调整为 224×224，并随机水平翻转，随机翻转概率为 0.5。

2.3　实验设置

本文采用随机梯度下降法(Stochastic Gradient Descent，SGD)训练，将初始学习率设为 0.01，动量为 0.9，权重衰减为 1×10^{-4}，在 RAF-DB 和 AffectNet 数据集上均迭代 120 次。由于不同数据集样本的差异性，在 RAF-DB 数据集上的批处理大小为 32，每 30 轮学习率以 0.1 的衰减率进行衰减。在 AffectNet 数据集上批处理大小为 64，每 10 轮学习率以 0.1

的衰减率衰减，此外，AffectNet 数据集的训练集是不平衡的，但测试集是平衡的，因此使用了均衡采样策略。

2.4 实验结果与分析

为了验证本文所提方法的有效性并衡量模型的计算复杂度，以 ShuffleNet-V2 作为主干网络，改进其输出层并增加 PDWRes 模块，引入标签分布学习，在大规模人脸表情数据集 RAF-DB 和 AffectNet 上进行实验，并就识别准确率与计算复杂度同其他方法进行了比较。

2.4.1 平衡系数 β 对分类效果的影响

该实验是为了探究 Combo Loss 损失函数中平衡系数 β 对不同人脸表情数据集识别准确率的影响。平衡系数 α 控制着 Dice Loss 对 $\mathcal{L}_2$ 的权重，实验时，对 Combo Loss 与改进交叉熵取平均分配相等的权重，即 $\alpha=0.5$ 。平衡系数 $\beta\in(0,1)$ 控制着改进交叉熵对 FP 和 FN 的惩罚程度，当 β 小于 0.5 时，由于 $(1-t_i)\ln(1-p_i)$ 的权重更大，FP 将比 FN 受到的惩罚更多，反之同理，实验时，β 以 0.1 的步幅从 0 到 1 进行取值。

在 RAF-DB 数据集上的实验结果如图 6 所示，表情识别准确率随着平衡系数 β 的递增先增加后下降，在 β 取 0.2 时，识别准确率达到最高 87.15%，当 β 小于 0.2 时，模型的识别准确率不足，当 β 的取值大于 0.2 时，模型的识别准确率开始下降。这表明对于 RAF-DB 数据集，需要对假阳性样本图片进行较大的惩罚，以辅助模型的学习取得较好的识别准确率。

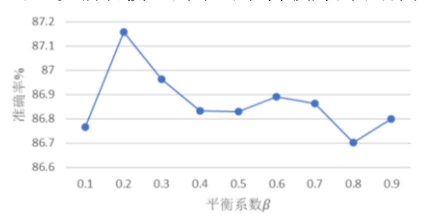


图 6 平衡系数 β 对 RAF-DB 数据集识别准确率影响

Fig. 6 The influence of balance coefficient β on the recognition accuracy of RAF-DB

在 AffectNet-7 数据集上的实验结果如图 7 所示，表情识别准确率随着平衡系数 β 的递增先下降后增加再下降，在 β 取 0.6 时，识别准确率达到最高 62.05%，当 β 小于 0.6 时，模型准确率先降后升，当 β 大于 0.6 时，模型的识别准确率开始明显下降。对于 AffectNet 数据集，需要对假阴性样本图片进行惩罚。实验结果表明平衡系数 β 对网络的识别效果有较大影响，不同数据集下平衡系数 β 的选择至关重要。

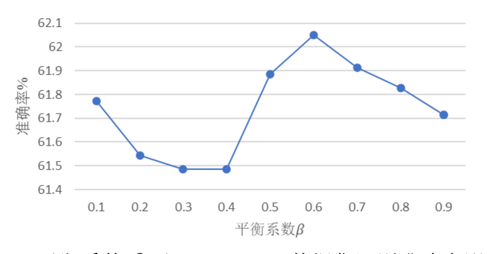


图 7 平衡系数 β 对 AffectNet-7 数据集识别准确率影响

Fig. 7 The influence of balance coefficient β on the recognition accuracy of affectnet-7

2.4.2 RAF-DB 实验结果

图 8 展示了在 RAF-DB 表情数据集上训练与测试的准确率曲线和损失函数曲线，为了在同一坐标系下清晰显示，将损失函数曲线放大 30 倍。从图中可以看出，在训练到第 35 轮时，模型已基本收敛，且最终训练集与测试集的识别准确率相差不大，这得益于改进 ShuffleNet 模型的输出模块，避免了模型的过拟合，模型最终在 RAF-DB 数据集上的识别准确率达到了 87.15%。

为了进一步验证本文模型的有效性，并衡量模型的计算复杂度，在 RAF-DB 表情数据集上与近年来其他文献的识别效果进行对比，如表 1 所示。在参数量方面，本文方法的参数量仅为 1.26M，这远低于 gACNN[14]方法的 134.29M，并且相较于参数量较小的 Separate Loss[15]、RAN[16]和 DDA Loss[18]方法，本文方法的参数量也仅为其十分之一，较大的压缩了模型的参数量。在浮点运算数方面，本文方法的 FLOPs 为 294.60M，相较于 gACNN 方法压缩了 98.09%的计算量，相较于 Separate Loss 和 DDA Loss 方法也压缩了 83.81%的计算量，使本文模型具有较低的复杂度。在准确率方面，相较于近期提出的 RAN 和 DDA Loss 方法，本文的准确率提升了 0.25%，相较于 IPA2LT[13]、gACNN、Separate Loss 和 LDL-ALSG[17]，本文模型的识别准确率也分别提升了 0.38%、2.08%、0.77%和 1.62%。由于数据集中的标签可能存在标注错误，wang 等人[19]提出了自治愈网络 SCN，通过正则化排序和重标签等操作纠正网络的学习，在 RAF-DB 数据集上取得了 87.03%的准确率，本文的准确率与之相比提升了 0.12%，且本文的参数量和 FLOPs 分别压缩了 10 倍和 6 倍，验证了本文方法的有效性。可以看出，本文方法在保持较低参数量与计算量的前提下，同时具有较好的识别准确率，这有利于本文模型在实际生产中的应用。

表 1 RAF-DB 数据集上不同方法的准确率和计算复杂度比较

Tab. 1 Comparison of accuracy and computational complexity of different methods on the RAF-DB

方法	年份	参数量(M)	FLOPs(M)	准确率(%)
IPA2LT[13]	2018	23.52	4109.48	86.77
gACNN[14]	2019	134.29	15479.79	85.07
Separate Loss[15]	2019	11.18	1818.56	86.38
RAN[16]	2020	11.19	14548.45	86.90
LDL-ALSG[17]	2020	23.52	4109.48	85.53
DDA Loss[18]	2020	11.18	1818.56	86.90
SCN[19]	2020	11.18	1818.56	87.03
LFER-LDL	**2021**	**1.26**	**294.60**	**87.15**

2.4.3 AffectNet 实验结果

图 9 展示了 AffectNet-7 表情数据集上训练与测试的准确率曲线和损失函数曲线，与 RAF-DB 实验一样，将损失函数曲线作同样的放大的处理。从图中可以看出，在训练到第 15 轮时，模型已基本收敛，具有较快的拟合速度，这得益于 LDL 模块辅助模型的学习，可以快速稳定的收敛，这也有利于模型在实际嵌入式设备上的运行，同样也避免了模型的过拟合问题，模型最终在大规模表情数据集 AffectNet-7 上的识别准确率达到了 62.05%。

为了验证本文方法在 AffectNet-7 数据集上的有效性及其计算复杂度，与近年来其他方法进行对比，对比情况如表 2。在参数量方面，本文方法仅为 VGG Face[21]方法参数量 145M 的 0.8%，相较于其他方法，本文方法的参数量仅为其 0.93%~51.01%，本文的参数量保持在较低水平。在浮点运算数方面，相较于 VGG Face 方法的 15490.46M，本文方法压缩了 98.1%的计算量，该压缩量与 gACNN 方法相比近似，相较于其他 7 种方法，本文方法也压缩了 61.40%~94.8%的计算量。在准确率方面，相较于近期提出的 VGG Face 和 LDL-ALSG，本文的识别准确率分别提升了 2.05%和 2.70%，相较于 IPA2LT、gACNN、Separate Loss、IPFR 和 FMPN 方法，本文的准确率也分别提升了 4.74%、3.27%、3.16%、4.65%和 0.53%，尽管本文的识别准确率不及 SNA-DFER 和 DDA Loss，但在参数量上 SNA-DFER 和 DDA Loss 方法分别为本文的 1.9 倍和 8.8 倍，在 FLOPs 上两者分别为本文的 2.5 倍和 6.1 倍，这不利于模型在性能较低的嵌入式设备上运行。综合来看，本文方法在有效降低模型复杂度的同时能保持较高水平的表情识别效果，验证了本文方法的有效性与实用性。

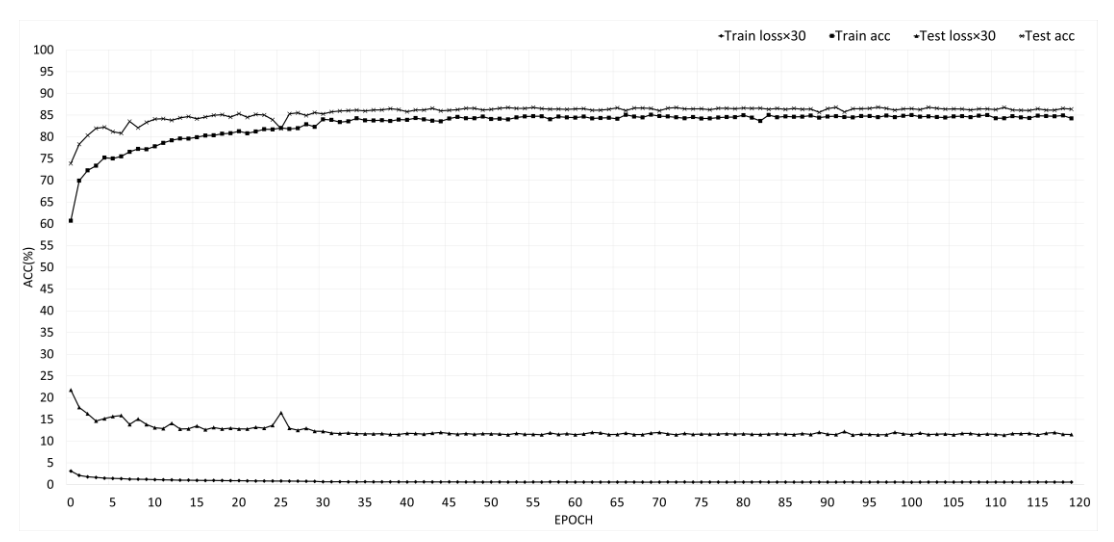


图 8　RAF-DB 的训练与测试曲线

Fig. 8　Training and testing curves of RAF-DB

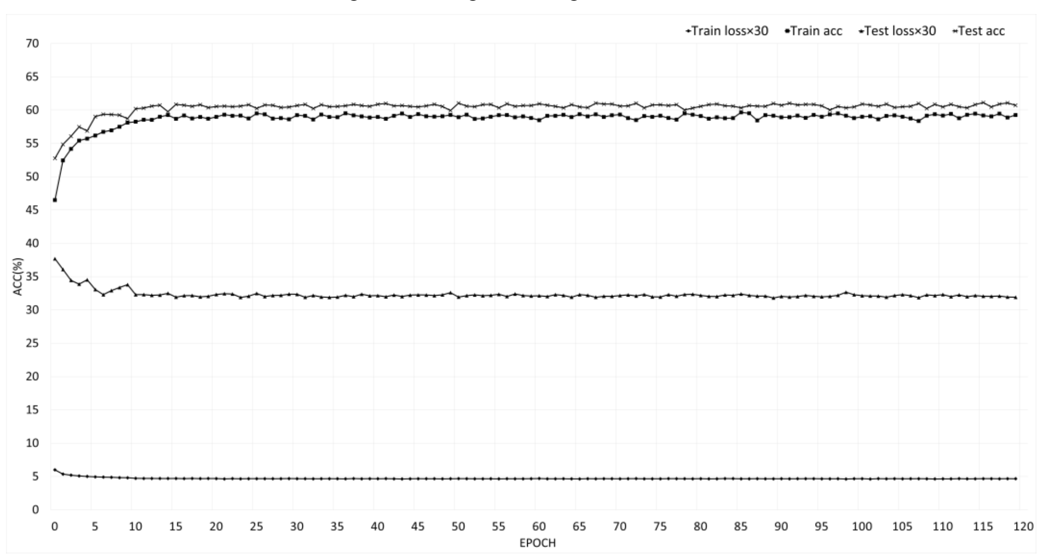


图 9　AffectNet-7 的训练与测试曲线

Fig. 9　Training and testing curves of affectnet-7

表 2　AffectNet-7 数据集上不同方法的准确率和计算复杂度比较

Tab. 2　Comparison of accuracy and computational complexity of different methods on the AffectNet-7

方法	年份	参数量(M)	FLOPs(M)	准确率(%)
IPA2LT[13]	2018	23.52	4109.48	57.31
gACNN[14]	2019	134.29	15479.79	58.78
IPFR[20]	2019	21.80	5729.12	57.40
Separate Loss[15]	2019	11.18	1818.56	58.89
FMPN[22]	2019	21.80	5729.12	61.52
VGG Face[21]	2020	145.00	15490.46	60.00
LDL-ALSG[17]	2020	23.52	4109.48	59.35
SNA-DFER[23]	2020	2.47	763.09	62.70
DDA Loss[18]	2020	11.18	1818.56	62.34
LFER-LDL	**2021**	**1.26**	**294.60**	**62.05**

同时，为了进一步验证本文方法在含有 8 类情感标签数据集 AffectNet-8 上的有效性，并评估其参数量和 FLOPs，与其他方法进行了对比分析，如表 3 所示，本文在 AffectNet-8 数据集上取得了 58.49%的准确率。在参数量方面，相较于 Weighted-loss、VGGNet-Variant 和 RAN 方法，本文的参数量仅为其 2.21%、19.27%和 11.26%。在浮点运算数方面，Weighted-loss 的 FLPOPs 约为本文方法的 2400 倍，RAN 和 ESR-9 方法分别为本文的 49 倍和 3 倍。在准确率方面，相较于 Weighted-loss、MobileNet-Variant 和 VGGNet-Variant 方法，本文提升了 0.49%、2.49%和 0.49%的准确率，尽管本文的准确率不及 RAN 和 ESR-9，但是本文的 FLOPs 都远低于二者。MobileNet-Variant 方法虽然在参数量和 FLOPs 上均取得了较好的效果，但是其准确率比本文低了约 2.5%。VGGNet-Variant 方法也实现了较低的 FLOPs，但是其在参数量和准确率上的表现不及本文，可以看出模型的计算复杂度和模型的性能两者不可兼得，本文在保持较低参数量和 FLOPs 的前提下，在 AffectNet-8 数据集上仍取得了不错的表现。

表 3　AffectNet-8 数据集上不同方法的准确率和计算复杂度比较

Tab. 3　Comparison of accuracy and computational complexity of different methods on the AffectNet-8

方法	年份	参数量(M)	FLOPs(M)	准确率(%)
Weighted-loss[7]	2017	57.03	710624.57	58.00
MobileNet-Variant[24]	2018	0.074	13.56	56.00
VGGNet-Variant[24]	2018	6.54	80.44	58.00
RAN[16]	2020	11.19	14548.45	59.50
ESR-9[25]	2020	0.37	1164.43	59.30
LFER-LDL	**2021**	**1.26**	**294.60**	**58.49**

从表 3 可以看出，AffectNet-8 是一个具有挑战性的人脸表情数据集，本文方法在 AffectNet-7 和 AffectNet-8 数据集上的准确率也有一定差异，AffectNet-8 在 AffectNet-7 的基础上增加了蔑视的表情，通过观察数据集发现，AffectNet-8 数据集的蔑视表情中存在大量非本表情的图像，例如快乐等，

由标注者的主观性造成的标签噪声，这将不利于网络的学习，图 10 展示了蔑视表情中的部分并不属于蔑视的图像。

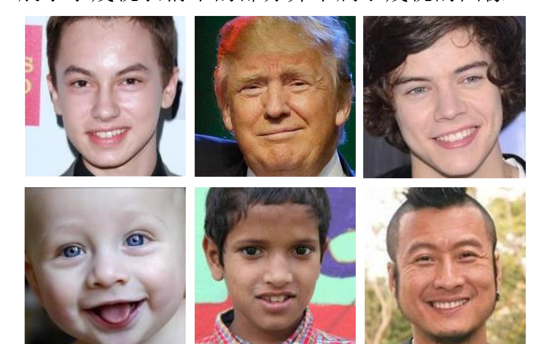


图 10　蔑视表情中包含的其他类别表情图像

Fig. 10　Other expression images in the contempt expression

2.4.4 消融实验

本文方法包括对输出模块的改进、并行深度卷积残差模块的设计以及标签分布学习，为了分析不同部分对人脸表情识别效果的影响，以 RAF-DB 数据集为例进行了消融实验。

本节以 ShuffleNet 为基线，依次加入改进的输出模块、并行深度卷积残差模块和标签分布学习，分析三个模块对识别性能的影响，实验结果如表 4 所示。通过改进输出模块，提取人脸表情高维特征，在参数量增加 0.02M 和 FLOPs 增加 0.03M 的情况下，识别准确率相较于 ShuffleNet 基线网络提升了 0.47%，这得益于改进输出模块中深度可分离卷积对人脸表情特征的进一步提取，同时使用的 Mish 激活函数也保证了特征信息的流动。并行深度卷积残差模块通过集成深度卷积获取局部区域特征，使网络更加关注不同表情中的细微差异，在参数量增加 0.01M 和 FLOPs 增加 3.09M 的情况下，相较于基线网络有 1.36%的提升，通过将局部特征融合到全局特征中，这使得模型更加关注人脸表情图像中具有鉴别性的特征，而这一特点与人眼的工作原理相似。标签分布学习有利于减少歧义表情的影响，而这并不会引入额外的参数量和 FLOPs，最终达到了 87.15%的准确率，相较于原始网络提升了 5.13%，现实世界中的人脸表情图像往往具有复杂的情感意图，标签分布学习通过收集数据集中表情图像的分布，来减少歧义表情的不确定性，这有利于缓解单标签所带来信息量不足的问题，说明了本文方法的有效性。

表 4　消融实验对比结果

Tab. 4　Ablation experiment comparison results

Baseline	改进输出模块	并行深度卷积残差模块	标签分布学习	参数量(M)	FLOPs(M)	准确率(%)
+	-	-	-	1.23	291.48	82.02
+	+	-	-	1.25	291.51	82.49
+	+	+	-	1.26	294.60	83.85
+	+	+	+	1.26	294.60	87.15

最后，为了将本文方法的结果进行可视化，将训练好的网络模型保存并进行人脸表情识别，在网上随机选取图像以及部分数据集用作实例测试，测试结果如图 11 所示。

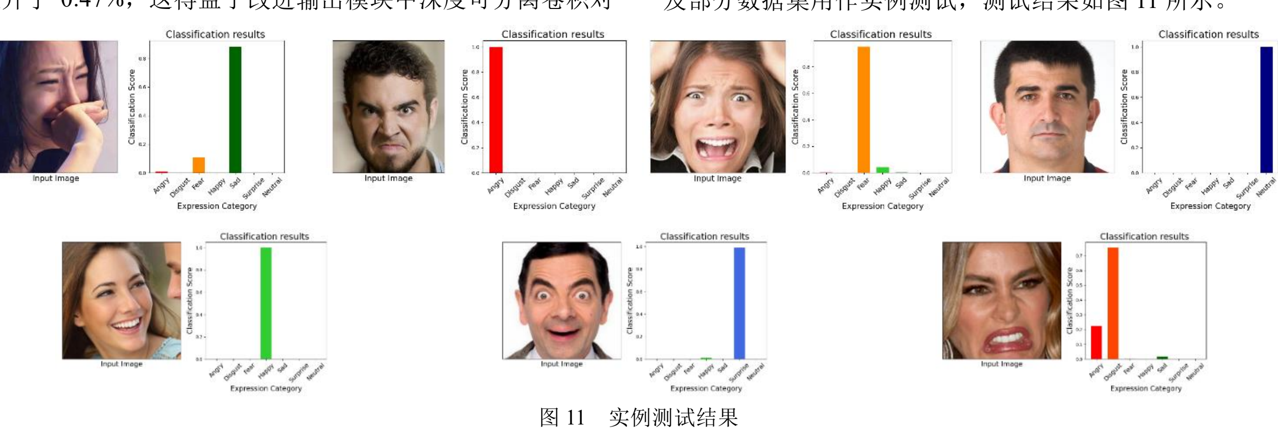


图 11　实例测试结果

Fig. 11　Example test results

3　结束语

人脸表情识别在诸多领域有着广泛的应用，但在实际生产环境中，过于复杂的网络模型不利于在配置受限的设备上运行。因此，本文提出了一种基于标签分布学习的轻量级人脸表情识别方法。本方法从特征提取的角度，对传统的 ShuffleNet 网络模型作出改进，并设计了并行深度卷积残差模块，这有利于增强模型对人脸表情图像中局部细节的特征提取能力；在训练策略上，通过标签分布学习，解决单标签信息量不足带来的歧义表情问题。最后，研究分析了 Combo Loss 损失函数中平衡系数对不同人脸表情数据集的影响。本文分别在 RAF-DB、AffectNet-7 和 AffectNet-8 数据集上做了对比实验，实验结果表明，本方法在保持较低参数量和 FLOPs 的前提下，仍具有较高的识别精度，具备较强的实用性。

深度学习模型在人脸表情识别研究中往往需要大量的标注数据，这不仅会产生昂贵的标注成本，而且可能会引入主观因素的标签噪声。因此在接下来的工作中，将研究如何进行半监督或无监督学习的人脸表情识别。

参考文献：

[1] Yu M, Guo Z, Yu Y, *et al.* Spatiotemporal featuredescriptor for micro-expression recognition using local cube binarypattern [J]. IEEE Access, 2019, 7: 214-225.

[2] 郑剑, 郑炽, 刘豪, 等. 融合局部特征与两阶段注意力权重学习的面部表情识别 [J]. 计算机应用研究, 2021. (Zheng Jian, Zheng Chi, Liu Hao, *et al.* Deep convolutional neural network fusing local feature and two-stage attention weight learning for facial expression recognition [J]. Application Research of Computers, 2021.)

[3] Ekman P, Friesen W V. Facial Action Coding System (FACS): A Technique for the Measurement of Facial Actions [J]. Rivista Di Psichiatria, 1978, 47 (2): 126-38.

[4] Plutchik R. A general psychoevolutionary theory of emotion [M]. Theories of emotion. Academic press, 1980: 3-33.

[5] Chen S, Wang J, Chen Y, *et al.* Label distribution learning on auxiliary label space graphs for facial expression recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13984-13993.

[6] Li S, Deng W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition [J]. IEEE Trans on Image Processing, 2019, 28 (1): 356-370.

[7] Ali M, Behzad H, Mohammad H. Mahoor. AffectNet: A New Database for Facial Expression, Valence, and Arousal Computation in the Wild [J].

IEEE Trans on Affective Computing, 2017.

[8] Ma N, Zhang X, Zheng H T, *et al.* Shufflenet v2: Practical guidelines for efficient cnn architecture design [C]// Proceedings of the European conference on computer vision (ECCV) . 2018: 116-131.

[9] Lin M, Chen Q, Yan S. Network in network [J]. arXiv preprint arXiv: 1312. 4400, 2013.

[10] Sifre L, Mallat S. Rotation, scaling and deformation invariant scattering for texture discrimination [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 1233-1240.

[11] Liu Z, Lin Y, Cao Y, *et al.* Swin transformer: Hierarchical vision transformer using shifted windows [J]. arXiv preprint arXiv: 2103. 14030, 2021.

[12] Deng J, Guo J, Ververas E, *et al.* Retinaface: Single-shot multi-level face localisation in the wild [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5203-5212.

[13] Zeng J, Shan S, Chen X. Facial expression recognition with inconsistently annotated datasets [C]// Proceedings of the European conference on computer vision (ECCV) . 2018: 222-237.

[14] Li Y, Zeng J, Shan S, *et al.* Occlusion aware facial expression recognition using CNN with attention mechanism [J]. IEEE Trans on Image Processing, 2018: 1-1.

[15] Li Y, Lu Y, Li J, *et al.* Separate loss for basic and compound facial expression recognition in the wild [C]// Asian Conference on Machine Learning. PMLR, 2019: 897-911.

[16] Wang K, Peng X, Yang J, *et al.* Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition [J]. IEEE Trans on Image Processing, 2020, PP (99): 1-1.

[17] Chen S, Wang J, Chen Y, *et al.* Label distribution learning on auxiliary label space graphs for facial expression recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13984-13993.

[18] Farzaneh A H, Qi X. Discriminant distribution-agnostic loss for facial expression recognition in the wild [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 406-407.

[19] Wang K, Peng X, Yang J, *et al.* Suppressing uncertainties for large-scale facial expression recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6897-6906.

[20] Wang C, Wang S, Liang G. Identity-and pose-robust facial expression recognition through adversarial feature learning [C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019: 238-246.

[21] Kollias D, Cheng S, Ververas E, *et al.* Deep neural network augmentation: Generating faces for affect analysis [J]. arXiv preprint arXiv: 1811. 05027, 2018.

[22] Chen Y, Wang J, Chen S, *et al.* Facial motion prior networks for facial expression recognition [C]// 2019 IEEE Visual Communications and Image Processing (VCIP) . IEEE, 2019: 1-4.

[23] Fu Y, Wu X, Li X, *et al.* Semantic neighborhood-aware deep facial expression recognition [J]. IEEE Transactions on Image Processing, 2020, 29: 6535-6548.

[24] Hewitt C, Gunes H. Cnn-based facial affect analysis on mobile devices [J]. arXiv preprint arXiv: 1807. 08775, 2018.

[25] Siqueira H, Magg S, Wermter S. Efficient facial feature learning with wide ensemble-based convolutional neural networks [C]// Proceedings of the AAAI conference on artificial intelligence. 2020, 34 (04): 5800-5809.